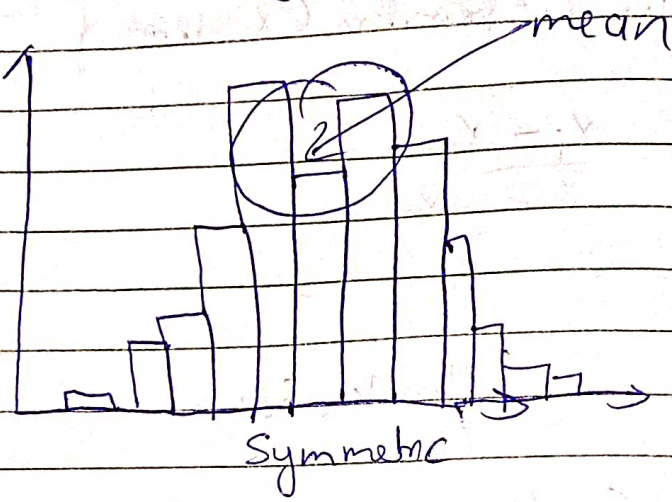
Date / /
Page No.

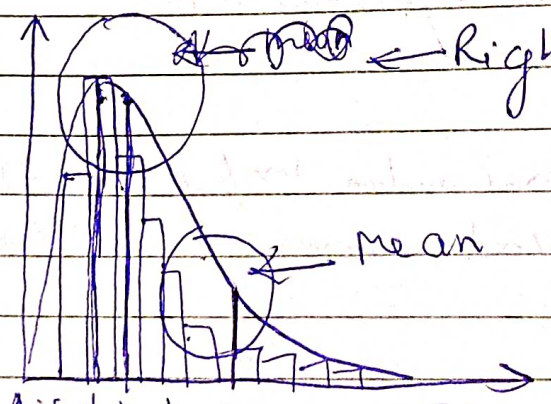
Measure of Central Tend

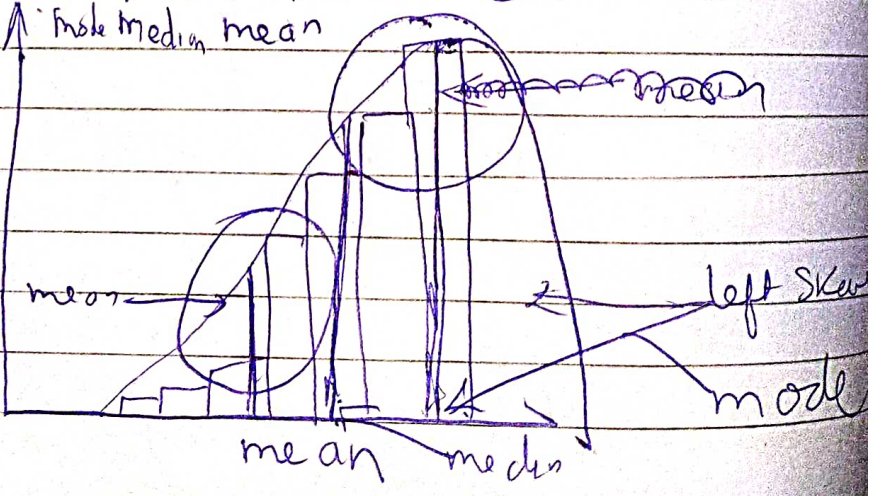
⇒ Centre point or typical value of a dataset

⇒ mean, median, mode indicate where values in a distribution fall and are referred to as the central location of a

① Mean — by Histogram







Formula —

$$\bar{x} = \frac{\sum f_i x_i}{\sum f_i}$$

(II) Median →

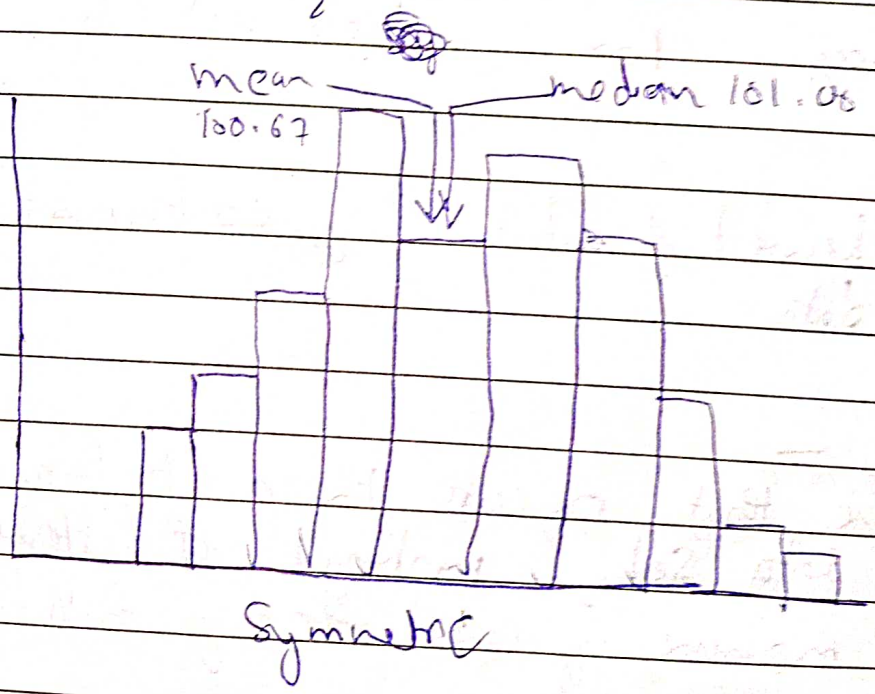
=> It is middle value

=> Median of odd dataset

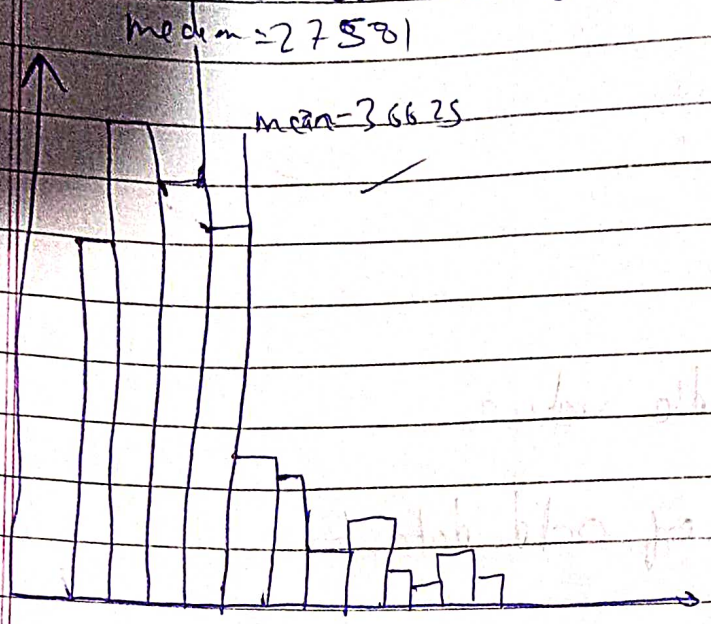
median = $\left(\frac{n+1}{2}\right)^{th}$ positive

=> Median of even dataset

median = $\frac{\left(\frac{n}{2}\right)^{th} + \left(\frac{n}{2}+1\right)^{th} \text{ post}}{2}$



Skewed continuous —



⇒ In skewed distribution, the outlier in the tail pull the mean away from center towards the longer tail

⇒ mean vs median differ by over 9000.

⇒ median better represents the central tendency for skewed distribution

Note: Skewed distribution, ⊗ Continuous data, Ordinal data

(iii) Mode —

⇒ value that occurs the most frequently in your data set, making it a different type of measure of central tendency than the mean or median

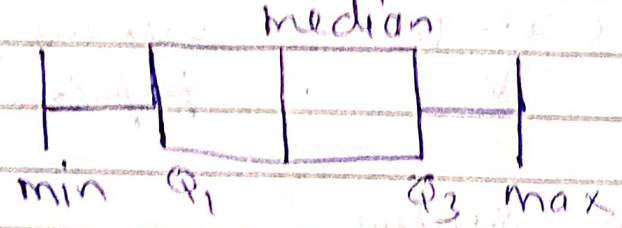
Measures of Dispersion

=> How spread out the values are
we measure "spread" using range, interquartile range, variance & standard deviation

Range —

=> Range is difference between the largest and smallest value in dataset

Range = df['Age'].max() - df['Age'].min()



Interquartile Range —

=> Difference between the first quartile and third quartile in a dataset.

$$IQR = Q_3 - Q_1$$

~~max~~ ~~df~~

Q_3 = df['Age'].percentile(.75)
Q_1 = df['Age'].percentile(.25)

Date / /
Page No.

Interquartile Range vs Range —

⇒ Interquartile range more resistant to outliers compared to the range which can make it a better metric to use mean "spread".

Variance —

It is a common way to measure how spread out data values are

$$\sigma^2 = \frac{\sum (x_i - \mu)^2}{N}$$ ← variance of population

Variance of sample

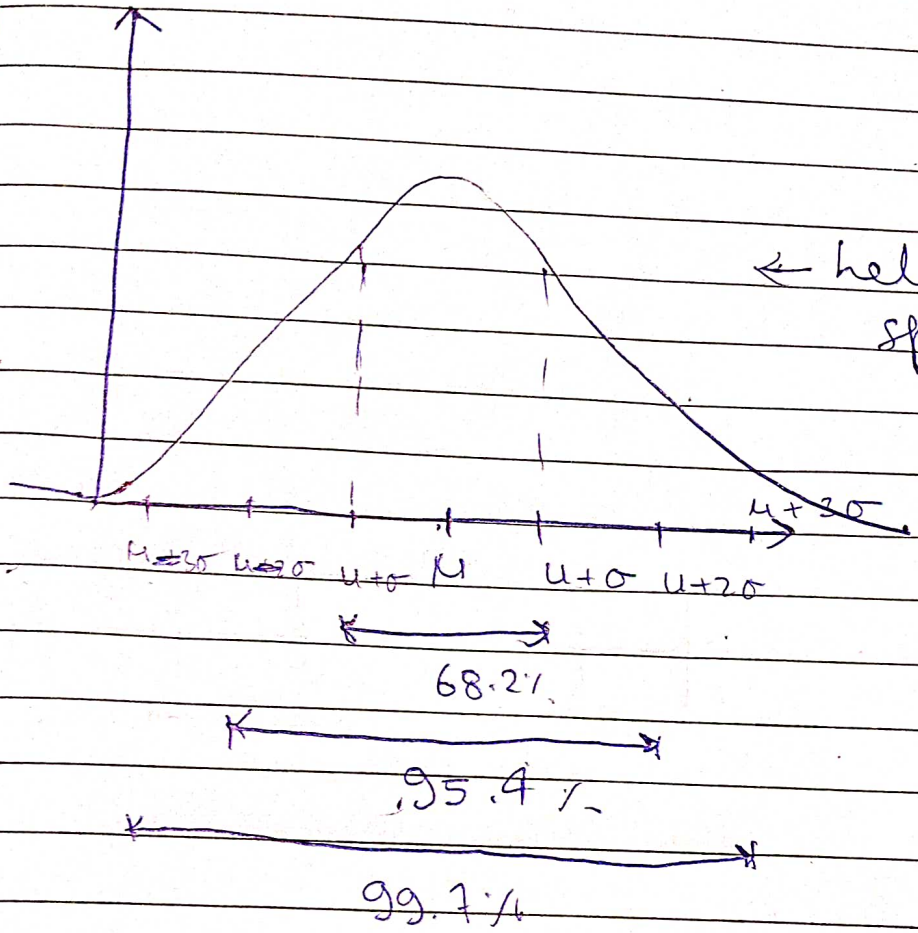
$$s^2 = \frac{\sum (x_i - \mu)^2}{n-1}$$

Standard Deviation —

Standard deviation is the square root of the variance. It's the most common way to measure how "spread out" data values are.

$$\sigma = \sqrt{\frac{\sum (x_i - \mu)^2}{N}}$$ ← standard deviation of population

$$s = \sqrt{\frac{\Sigma(x_i - \bar{x})^2}{n-1}}$$ ← Standard deviation of a sample



← helps to understand spread of dates